

Docket: 13020-10

## AUTOMATIC GENOTYPE DETERMINATION

03648178 071800

Inventors: Stephen E. Lincoln  
Michael R. Knapp

**Certificate of Express Mail Under 37 CFR 1.10**

I hereby declare that this correspondence is being deposited with the United States Postal Service via Express Mail Label No. EL301953995US in an envelope addressed to: Commissioner of Patents and Trademarks, Washington, DC 20231

July 18, 2000  
Date

Evan McHale  
Name

Attorney Docket 1471/108

## AUTOMATIC GENOTYPE DETERMINATION

Cross Reference to Related Applications

This application is <sup>a</sup> a continuation of application  
 serial no. 08/362,266, filed December 22, 1994, which is a  
 continuation in part of application serial no. 08/173,173,  
 filed December 23, 1993, which is for an invention entitled  
 "Automatic Genotype Determination," by Stephen E. Lincoln  
 and Michael P. Knapp. This immediate parent application <sup>a2</sup> is a  
 continuation in part of application serial no. 07/775,786,  
 filed October 11, 1991, for an invention entitled "Nucleic  
 Acid Typing by Polymerase Extension of Oligonucleotides  
 using Terminator Mixtures," by P. Goelet, M. Knapp, and S.  
 Anderson, which in turn is a continuation in part of  
 application serial no. 07/664,837, filed March 5, 1991.  
 Immediate parent application serial no. 08/173,173 is also a  
 continuation in part of application serial no. 08/162,397,  
 filed December 6, 1993, for an invention entitled "Method  
 for Immobilization of Nucleic Acid Molecules" by T.  
 Nikiforov and M. Knapp, and of application serial no.  
 08/155,746, filed November 23, 1993, for an invention  
 entitled "Method for Generating Single-Stranded DNA  
 Molecules" by T. Nikiforov and M. Knapp, and of application  
 serial no. 08/145,145, filed November 3, 1993, for an  
 invention entitled "Single Nucleotide Polymorphisms and  
 their use in Genetic Analysis" by M. Knapp and P. Goelet.  
 All of these related applications are hereby incorporated  
 herein by reference.

5  
AmoA110  
AmoA2
 10  
 08/173,173  
 07/775,786  
 07/664,837  
 08/162,397  
 08/155,746  
 08/145,145

20

25

5

Technical Field

The present invention relates to the methods and devices for determining the genotype at a locus within genetic material.

10

Summary of the Invention

05415  
05416  
05417  
05418  
05419  
05420  
05421  
05422  
05423  
05424  
05425  
05426  
05427  
05428  
05429  
05430  
05431  
05432  
05433  
05434  
05435  
05436  
05437  
05438  
05439  
05440  
05441  
05442  
05443  
05444  
05445  
05446  
05447  
05448  
05449  
05450  
05451  
05452  
05453  
05454  
05455  
05456  
05457  
05458  
05459  
05460  
05461  
05462  
05463  
05464  
05465  
05466  
05467  
05468  
05469  
05470  
05471  
05472  
05473  
05474  
05475  
05476  
05477  
05478  
05479  
05480  
05481  
05482  
05483  
05484  
05485  
05486  
05487  
05488  
05489  
05490  
05491  
05492  
05493  
05494  
05495  
05496  
05497  
05498  
05499  
05500  
05501  
05502  
05503  
05504  
05505  
05506  
05507  
05508  
05509  
05510  
05511  
05512  
05513  
05514  
05515  
05516  
05517  
05518  
05519  
05520  
05521  
05522  
05523  
05524  
05525  
05526  
05527  
05528  
05529  
05530  
05531  
05532  
05533  
05534  
05535  
05536  
05537  
05538  
05539  
05540  
05541  
05542  
05543  
05544  
05545  
05546  
05547  
05548  
05549  
05550  
05551  
05552  
05553  
05554  
05555  
05556  
05557  
05558  
05559  
05560  
05561  
05562  
05563  
05564  
05565  
05566  
05567  
05568  
05569  
05570  
05571  
05572  
05573  
05574  
05575  
05576  
05577  
05578  
05579  
05580  
05581  
05582  
05583  
05584  
05585  
05586  
05587  
05588  
05589  
05590  
05591  
05592  
05593  
05594  
05595  
05596  
05597  
05598  
05599  
05600  
05601  
05602  
05603  
05604  
05605  
05606  
05607  
05608  
05609  
05610  
05611  
05612  
05613  
05614  
05615  
05616  
05617  
05618  
05619  
05620  
05621  
05622  
05623  
05624  
05625  
05626  
05627  
05628  
05629  
05630  
05631  
05632  
05633  
05634  
05635  
05636  
05637  
05638  
05639  
05640  
05641  
05642  
05643  
05644  
05645  
05646  
05647  
05648  
05649  
05650  
05651  
05652  
05653  
05654  
05655  
05656  
05657  
05658  
05659  
05660  
05661  
05662  
05663  
05664  
05665  
05666  
05667  
05668  
05669  
05670  
05671  
05672  
05673  
05674  
05675  
05676  
05677  
05678  
05679  
05680  
05681  
05682  
05683  
05684  
05685  
05686  
05687  
05688  
05689  
05690  
05691  
05692  
05693  
05694  
05695  
05696  
05697  
05698  
05699  
05700  
05701  
05702  
05703  
05704  
05705  
05706  
05707  
05708  
05709  
05710  
05711  
05712  
05713  
05714  
05715  
05716  
05717  
05718  
05719  
05720  
05721  
05722  
05723  
05724  
05725  
05726  
05727  
05728  
05729  
05730  
05731  
05732  
05733  
05734  
05735  
05736  
05737  
05738  
05739  
05740  
05741  
05742  
05743  
05744  
05745  
05746  
05747  
05748  
05749  
05750  
05751  
05752  
05753  
05754  
05755  
05756  
05757  
05758  
05759  
05760  
05761  
05762  
05763  
05764  
05765  
05766  
05767  
05768  
05769  
05770  
05771  
05772  
05773  
05774  
05775  
05776  
05777  
05778  
05779  
05780  
05781  
05782  
05783  
05784  
05785  
05786  
05787  
05788  
05789  
05790  
05791  
05792  
05793  
05794  
05795  
05796  
05797  
05798  
05799  
05800  
05801  
05802  
05803  
05804  
05805  
05806  
05807  
05808  
05809  
05810  
05811  
05812  
05813  
05814  
05815  
05816  
05817  
05818  
05819  
05820  
05821  
05822  
05823  
05824  
05825  
05826  
05827  
05828  
05829  
05830  
05831  
05832  
05833  
05834  
05835  
05836  
05837  
05838  
05839  
05840  
05841  
05842  
05843  
05844  
05845  
05846  
05847  
05848  
05849  
05850  
05851  
05852  
05853  
05854  
05855  
05856  
05857  
05858  
05859  
05860  
05861  
05862  
05863  
05864  
05865  
05866  
05867  
05868  
05869  
05870  
05871  
05872  
05873  
05874  
05875  
05876  
05877  
05878  
05879  
05880  
05881  
05882  
05883  
05884  
05885  
05886  
05887  
05888  
05889  
05890  
05891  
05892  
05893  
05894  
05895  
05896  
05897  
05898  
05899  
05900  
05901  
05902  
05903  
05904  
05905  
05906  
05907  
05908  
05909  
05910  
05911  
05912  
05913  
05914  
05915  
05916  
05917  
05918  
05919  
05920  
05921  
05922  
05923  
05924  
05925  
05926  
05927  
05928  
05929  
05930  
05931  
05932  
05933  
05934  
05935  
05936  
05937  
05938  
05939  
05940  
05941  
05942  
05943  
05944  
05945  
05946  
05947  
05948  
05949  
05950  
05951  
05952  
05953  
05954  
05955  
05956  
05957  
05958  
05959  
05960  
05961  
05962  
05963  
05964  
05965  
05966  
05967  
05968  
05969  
05970  
05971  
05972  
05973  
05974  
05975  
05976  
05977  
05978  
05979  
05980  
05981  
05982  
05983  
05984  
05985  
05986  
05987  
05988  
05989  
05990  
05991  
05992  
05993  
05994  
05995  
05996  
05997  
05998  
05999  
06000

The present invention provides in one embodiment a method of determining the genotype at a locus within genetic material obtained from a biological sample. In accordance with this method, the material is reacted at the locus to produce a first reaction value indicative of the presence of a given allele at the locus. There is formed a data set including the first reaction value. There is also established a set of one or more probability distributions; these distributions associate hypothetical reaction values with corresponding probabilities for each genotype of interest at the locus. The first reaction value is applied to each probability distribution to determine a measure of the conditional probability of each genotype of interest at the locus. The genotype is then determined based on these measures.

25

In accordance with a further embodiment of this method, the material at the locus is subject to a second reaction to produce a second reaction value independently indicative of the presence of a second allele at the locus. A second data set is formed and the second reaction value is included in the second data set. Each probability distribution associates a hypothetical pair of first and second reaction

30

5 values with a single probability of each genotype of  
interest. The first data set includes other reaction values  
obtained under conditions comparable to those under which  
the first reaction value was produced, and the second data  
set includes other reaction values obtained under conditions  
10 comparable to those under which the second reaction value  
was produced. Where, for example, there are two alleles of  
interest, the first reaction may be an assay for one allele  
and the second reaction may be a distinct assay for the  
other allele. The first and second data sets may include  
reaction values for the first and second reactions  
15 respectively, run under comparable conditions on other  
samples with respect to the same locus. Alternatively, or in  
addition, the data sets may include reaction values for  
reactions run under comparable conditions with respect to  
different loci within the same sample.  
20

In accordance with a further embodiment, the  
probability distributions may be determined iteratively. In  
this embodiment, each probability distribution is initially  
estimated. Each initial probability distribution is used to  
25 determine initial genotype probabilities using the reaction  
values in the data sets. The resulting data are then used to  
modify the initial probability distribution, so that the  
modified distribution more accurately reflects the reaction  
values in the data set. This procedure may be iterated a  
30 desired number of times to improve the probability  
distribution. In practice, we have generally found that a  
single iteration is sufficient.

05:41:51.78 "072000

5       The foregoing methods have been employed with success  
for automatic genotype determination based on assays using  
genetic bit analysis (GBA). In such a case, each allele may  
typically be a single specific nucleotide. In accordance  
with GBA, a reaction is designed to produce a value that is  
10       indicative of the presence of a specific allele at the locus  
within the genetic material. In GBA, the approach is  
typically to hybridize a specific oligonucleotide to the  
genetic material at the locus immediately adjacent to the  
nucleotide being interrogated. Next, DNA polymerase is  
15       applied in the presence of differentially labelled  
dideoxynucleoside triphosphates. The read-out steps detect  
the presence of one or more of the labels which have become  
covalently attached to the 3' end of the oligonucleotide.  
Details are provided in Theo R. Nikiforov et al. "Genetic  
20       Bit Analysis, a solid phase method for typing single  
nucleotide polymorphisms," 22 Nucleic Acids Research, No.  
20, 4167-4175 (1994), which is hereby incorporated herein by  
reference. However, the present invention is also applicable  
to other reaction systems for allele determination, such as  
25       allele-specific hybridization (ASH), sequencing by  
hybridization (CBH), oligonucleotide ligase assay (OLA), and  
allele-specific amplification, using either the ligase chain  
reaction (LCR) or the polymerase chain reactions (PCR). The  
alleles assayed may be defined, for example, by a single  
30       nucleotide, a pair of nucleotides, a restriction site, or  
(at least in part) by its length in nucleotides.

In another embodiment of the invention, there is

5 at the locus. Thereafter the distribution establishment arrangement modifies each initial probability distribution, so that each modified distribution more accurately reflects the reaction values stored in the storage means.

10 The term "reaction value" as used in this description and the following claims may refer either to a single numerical value or to a collection of numbers associated with a physical state produced by the reaction. In the GBA method described in the Nikiforov article referred to above, e.g., optical signals are produced that may be read as a single numerical value. Alternatively, e.g., an optical signal may be simplified over time, and the reaction value may be the collection of samples of such a signal. It is also possible to form a scanned image, of one or a series of optical signals generated by GBA or other reaction methods, and to digitize this image, so that a collection of pixel values in all or a portion of the image constitutes a reaction value.

#### Brief Description of the Drawings

25 The foregoing aspects of the invention will be more readily understood by reference to the following detailed description, taken with respect to the following drawings, in which:

30 Fig. 1 is a diagram of a device in accordance with a preferred embodiment of the invention;

Fig. 2 is a diagram of the logical flow in accordance

5 obtained under comparable conditions. A distribution  
establishment arrangement establishes a set of probability  
distributions, including at least one distribution,  
associating hypothetical reaction values with corresponding  
probabilities for each genotype of interest at the locus. A  
10 genotype calculation arrangement applies the first reaction  
value to each pertinent probability distribution to  
determine the conditional probability of each genotype of  
interest at the locus. A genotype determination arrangement  
determines the genotype based on data from the genotype  
5 calculation arrangement.

In a further embodiment, the device may determine the  
genotype at selected loci. In this embodiment, the reaction  
generation arrangement can produce a reaction value  
indicative of the presence of a given allele at each of the  
selected loci and the data set includes reaction values  
10 obtained with respect to each of the selected loci. The  
genotype calculation arrangement applies reaction values  
obtained with respect to each of the selected loci to each  
pertinent probability distribution.

25 In another further embodiment, the device may determine  
the genotype at a locus within genetic material from each of  
a plurality of samples. In this embodiment, the reaction  
generation arrangement can produce a reaction value  
indicative of the presence of a given allele at the locus of  
30 material obtained from each sample and the data set includes  
reaction values obtained with respect to each sample. The  
genotype calculation arrangement applies reaction values

10

15

30



5 provided a method of determining the genotype of a subject  
by reacting genetic material taken from the subject at  
selected loci. In this embodiment, each locus may be an  
identified single nucleotide or group of nucleotides, and  
there is produced with respect to each of the selected loci  
10 a reaction value indicative of the presence of a given  
allele at each of the selected loci. These reaction values  
are used to determine the genotype of the subject or  
alternatively a DNA sequence associated with a specific  
region of genetic material of the subject. (Indeed a set of  
15 genotypes for selected proximal loci may be used to specify  
a sequence of the genetic material.). In further embodiments,  
the loci are selected to provide one or more types of  
information concerning the subject, including inheritance of  
a trait, parentage, identity, and matching tissue with that  
20 of a donor. Alternatively, the loci may be spaced  
throughout the entire genome of subject to assist in  
characterizing the genome of the species of the subject.

In a further embodiment of the invention, there is  
provided a device for determining the genotype at a locus  
25 within genetic material obtained from a subject. The device  
of this embodiment has a reaction value generation  
arrangement for producing a first physical state,  
quantifiable as a first reaction value, indicative of the  
presence of a given allele at the locus, the value  
30 associated with reaction of the material at the locus. The  
device also has a storage arrangement for storing a data set  
including the first reaction value and other reaction values

5 with the embodiment of Fig. 1;

Fig. 3 is a graph of numeric reaction values (data) generated by the embodiment of Fig. 1 as well as the genotype determinations made by the embodiment from these data; and

10 Figs. 4-7 show probability distributions derived by the embodiment of Fig. 1 for three genotypes of interest (AA, AT, and TT) and a failure mode at a locus.

Fig. 8 is an example of the output of the device in Fig. 1.

#### 15 Detailed Description of Specific Embodiments

The invention provides in preferred embodiments a method and device for genotype determination using genetic marker systems that produce allele-specific quantitative signals. An embodiment uses computer processing, employing computer software we developed and call "GetGenos", of data produced by a device we also developed to produce GBA data. The device achieves, among other things, the following:

- 25
- Fully automatic genotype determination from quantitative data. Off-line analysis of data pools is intended, although the software is fast enough to use interactively.
  - Ability to examine many allele tests per DNA sample simultaneously. One genotype and confidence measure are produced from these data.
- 30
- A true probabilistic confidence measure (a LOD

5 score), properly calibrated, is produced for each genotype.

- Use of robust statistical methods: Noise reduction via selective data pooling and simultaneous search over points in a data pool, preventing bias.

- Maximal avoidance of arbitrary parameters, and thus  
10 insensitivity to great variation in input data. The small number of parameters that are required by the underlying statistical model are fit to the observed data, essentially using the data set as its own internal control.

- Flexibility for handling multiple data types.  
15 Essentially, only probability distribution calculations, described below, need to be calibrated to new data types. We expect that the invention may be applied to GBA, OLA, ASH, and RAPD-type markers.

20 Our current embodiment of the software is implemented in portable ANSI C, for easy integration into a custom laboratory information system. This code has been successfully run on:

- Macintosh
- Sun
- 25 • MS-DOS
- MS-Windows

In our current embodiment of the software, a number of consistency checks are performed for GBA data verification, using both the raw GBA values and the control wells.

30 Overall statistics for trend analysis and QC are computed. Brief "Genotype Reports" are generated, summarizing results for each data set, including failures. All data are output

5 in a convenient form for import into interactive statistical packages, such as DataDesk™. The current implementation is presently restricted to 2-allele tests in diploids - the situation with present GBA applications.

Referring to Fig. 1, there is shown a preferred  
10 embodiment of a device in accordance with the present invention. The device includes an optical detector 11 to produce reaction values resulting from one or more reactions. These reactions assay for one or more alleles in samples of genetic material. We have implemented the  
15 detector 11 using bichromatic microplate reader model 348 and microplate stacker model 83 from ICN Biomedical, Inc., P.O. Box 5023, Costa Mesa, California 92626. The microplates are in a 96 well format, and the reader accommodates 20  
20 microplates in a single processing batch. Accordingly the device of this embodiment permits large batch processing. The reactions in our implementation use GBA, as described above. The detector 11 is controlled by computer 12 to cause selected readout of reaction values from each well. The computer 12 is programmed to allow for multiple readout of  
25 the reaction value from a given well over a period of time. The values are stored temporarily in memory and then saved in database 14. Computer 13 accesses the database 14 over line 15 and processes the data in accordance with the procedure described below. Of course, computers 12 and 13  
30 and database 14 may be implemented by an integral controller and data storage arrangement. Such an arrangement could in fact be located in the housing of the optical detector 11.

5 In Fig. 2 is shown the procedure followed by computer  
13. The steps of this procedure are as follows:

Input Data: A set of data is loaded under step 21. In  
most applications, each experiment in the set should be  
testing (i) the same genetic marker, and (ii) the same set  
10 of alleles of that marker, using comparable biochemistry  
(e.g. the same reagent batches, etc.). Large data sets help  
smooth out noise, although the appropriate size of a data  
set depends on the allele frequencies (and thus the number  
of expected individuals of each genotypic class). Each data  
15 point in the input data may be thought of as an N-tuple of  
numeric values, where N is the number of signals collected  
from each DNA sample for this locus. (N will usually be the  
number of alleles tested at this marker, denoted A, except  
when repeated testing is used, in which case N may be  
20 greater than A).

Preprocess Data: Next the data are subject to  
preprocessing (step 22). An internal M-dimensional Euclidean  
representation of the input signals is produced, where each  
input datum (an N-tuple) is a point in M-space. Usually, M  
25 will be the same as N and the coordinates of the point will  
be the values of the input tuple, and thus the preprocessing  
will be trivial (although see the first paragraph of  
variations discussed). The Euclidean space may be  
non-linear, depending on the best available models of signal  
30 generation. (Completely mathematically equivalently, any  
non-linearity may be embodied in the initial probability  
distributions, described below.)

Fig. 3 illustrates preprocessed reaction values from step 22 for GBA locus 177-2 on 80 DNA samples. The X-axis indicates preprocessed reaction values for allele 1 (A) and the Y-axis indicates preprocessed reaction values for allele 2 (T). For clarity, the results of genotype determination are also indicated for each point: Triangles are TT genotype, diamonds are AA, circles are AT, and squares are failures (no signal).

Probability Distributions: Returning to Fig. 2, under step 22, initial probability distributions are established for the G possible genotypes. For example, in a random diploid population containing A tested alleles:

$$G = (A) + (A - 1) + \dots + 1 = \frac{A(A + 1)}{2} \quad (1)$$

The initial conditional probability for any hypothetical input datum (a point in M-space, denoted  $X_i$ ) and genotype (denoted  $g$ ) is defined as the prior probability of seeing the signal  $X_i$  assuming that  $g$  is the correct genotype of that datum. That is:

$$\begin{aligned} &\Pr(\text{signal } X_i \cdot \text{Genotype} = g), \\ &\text{where } X_i = (x_1^1 \dots x_1^M) \text{ and } g \in \{1 \dots G\} \end{aligned} \quad (2)$$

Figures 4 through 7 illustrate the initial probability distributions established for the data in figure 3. Probability distributions are indicated for the four

5 genotypic classes of interest, AA, AT, TT and No Signal, in  
Figs 4, 5, 6, and 7 respectively. The shading at each XY  
position indicates probability, with darker shades  
indicating increased probability for hypothetical data  
points with those X and Y reaction valves.

10 Exactly where these distributions come from is highly  
specific to the nature of the input data. The probability  
distributions can either be pre-computed at this step and  
stored as quantized data, or can be calculated on the fly as  
needed in step 23, below. The probability distributions may  
be fixed, or may be fit to the observed data or may be fit  
to assumed genotypes as determined by previous iterations of  
this algorithm. (See Additional Features below.)

Under step 23, we compute the conditional probability  
of each genotype. For each datum  $X_i$ , the above probabilities  
are collected into an overall conditional posterior  
probability of each genotype for that datum:

$$\begin{aligned} &\Pr(\text{Genotype} = g \mid \text{Signal } X_i) = \\ &\frac{\Pr(\text{Signal } X_i) \mid \text{Genotype} = g) \cdot \Pr(\text{Genotype} = g)}{\Pr(\text{Signal } X_i)} \quad (3) \end{aligned}$$

where

$\Pr(\text{Genotype} = g)$  is the prior probability of any datum  
having genotype  $g$ ;

$\Pr(\text{Signal } X_i)$  is the prior probability of the signal (a  
constant which may be ignored); and

$\Pr(\text{Signal } X_i) \cdot \text{Genotype} = g)$  is the initial  
probability defined above.

5 Under step 24, we determine the select the genotype and  
 compute the confidence score. For each datum, using the  
 above posterior probabilities, we determine the most likely  
 genotype assignment  $g'$  (the genotype with the highest  
 posterior probability) and its confidence score. The  
 10 confidence score  $C$  is simply the log of the odds ratio:

$$C = \log_{10} \frac{\Pr(\text{Genotype} = g' \mid \text{Signal } X_i)}{\sum_{\text{Genotypes } g} \Pr(\text{Genotype} = g \mid \text{Signal } X_i)} \quad (4)$$

15 It should be noted that this procedure is significant, among  
 other reasons, because it permits determining a robust  
 probabilistic confidence score associated with each geno  
 type determination.

20 Under step 25, there may be employed adaptive fitting.  
 A classic iterative adaptive fitting algorithm, such as  
 Estimation-Maximization (E-M), may be used to increase the  
 ability to deal with highly different input data sets and  
 reduce noise sensitivity. In this case, the genotypes  
 25 computed in step 24 are used to refit the distributions  
 (from step 22). In step 25, a convergence test is performed,  
 which may cause the program to loop back to step 23, but now  
 using the new distributions.

30 As one example, an E-M search procedure may be used to  
 maximize the total likelihood, that is, to find the  
 maximally likely set of genotype assignments given the input  
 data set. (The net likelihood may be calculated from the  
 Bayesian probabilities, defined above.) For appropriate



5       likelihood calculations and probability distributions, the  
EM principle will guarantee that this algorithm always  
produces true-maximum-likelihood values, regardless of  
initial guess, and that it always converges.

10       Output Data: Under step 26, we output the results  
(genotypes and confidence scores) to the user or to a  
computer database. An example of such output is shown in  
Fig. 8.

#### Additional Features

15       Additional features may be incorporated into the above  
procedure. They may be integrated into the procedure either  
together or separately, and have all been implemented in a  
preferred embodiment.

20       Preprocessing: During steps 21 or 22, the data (either  
input tuples or spatial data points) may be preprocessed in  
order to reduce noise, using any one of many classical  
statistical or signal-processing techniques. Control data  
points may be used in this step. In fact, various types of  
signal filtering or normalizing may be applied at almost any  
step in the algorithm.

25       Fitting Probability Distributions: The probability  
distributions calculated in steps 22 and 23 may be fit to  
the input data - that is, each distribution may be a  
function of values which are in part calculated from the  
input data. For example, we may define the conditional  
30       probability of a signal point for some genotype to be a  
function of the distance between that point and the observed  
mean for that signal.

5           Using an Initial Genotype Guess: In step 22, either a  
 simple or heuristic algorithm may be used to produce an  
 initial genotype guess for each input data point. If a  
 fairly accurate guess can be produced, then the probability  
 distributions for each genotype may be fit to the subset of  
 10   the data assumed to be of that genotypic class. Another use  
 of a genotype guess is in initial input validity checks  
 and/or preprocessing (e.g. Step 22), before the remainder of  
 the algorithm is applied. To be useful, a guess need not  
 produce complete genotypic information, however.

15           Using a Null Genotypic Class: In steps 22 and all  
 further steps, one (or more) additional probability  
 distributions may be added to fit the data to the signals  
 one would expect to see if an experiment (e.g. that datum)  
 failed. E.g.,

$$\text{Pr}(\text{signal } X_i \cdot \text{Genotype} \cdot \{1 \dots G\})$$

20           The current implementation above is presently  
 restricted to  $M=2$  and  $N=2 \cdot R$ , where  $R$  is the number of  
 25   repeated tests of both alleles. We refer to the two alleles  
 as  $X$  and  $Y$ . The program understands the notion of "plates"  
 of data, a number of which make up a data set.

30           The Initial Guess Variation is employed to initially  
 fit distributions using the heuristic described below. The  
 Initial Guess is produced during the Preprocessing Step  
 which normalizes and background subtracts the input data,  
 and remove apparent outlier points as well. These steps are

5 performed separately for each allele's signal (i.e., 1  
dimensional analysis). In fact, this preprocessing is  
applied separately to each of the R repeated tests, and the  
test with the small total 2 dimension residual is chosen for  
10 use in further steps. Various other preprocessing and  
post-processing steps are employed for GBA data validation  
and QC. In particular, controls producing a known reaction  
value may be employed to assure integrity of the biochemical  
process. In a preferred embodiment, signals are assumed to  
be small positive numbers (between 0.0 and 5.0, with 0.0  
15 indicating that allele is likely not present in the sample,  
and larger values indicating that it may be.

To handle a wide range of input data signal strengths,  
the Adaptive Fitting Variation is employed. However, the  
program is hard-coded to perform exactly one or two  
20 interactions passes through step 25, which we find works  
well for existing GBA data.

The probability distributions we fit at present in  
steps 22 and 25 have as their only parameters (i) the ratio  
of the X and Y signals for heterozygotes, and (ii) the  
25 variance from the normalized means (0.0 negative for that  
allele, 1.0 for positive for that allele) along each axis  
separately. In fact, these later numbers are constrained to  
be at least a fixed minimum, which is rarely exceeded, so  
that the algorithm will work with very small quantities of  
30 data and will produce the behavior we want. These numbers  
are computed separately for each microtiter plate. The  
probability distributions are generated using the code

5 (written in C) attached hereto and incorporated herein by reference as Appendix A.

The Null-Class variant is used to provide genotypic class indicating No Signal.

10 Quality control may also be enhanced in a surprising manner using the procedures described here. In particular, the confidence score C of equation (4) serves as a robust indicator of the performance of the biochemical reaction system. For example, a downward trend in the confidence scores within a single batch or in successive batches may indicate deterioration of an important reagent or of a sample or miscalibration of the instrumentation.

15 Accordingly, in a preferred embodiment, the computer may be used to determine the presence of a downward trend in the confidence score over time calculated in reference to each of the following variables: the locus (is there a downward trend in the confidence score of a single locus relative to other loci tested?), the sample (is there a downward trend in the confidence score of a single sample relative to other samples tested?), plate (is there a downward trend in the confidence score of this plate relative to other plate?), and batch (relative to other batches). If a downward trend of statistical significance (using, for example a chi square test) is detected, an alarm condition is entered.

25 30 Because the confidence score is an accurate indication of the reliability of the reaction system and the genotype determination, a low confidence score associated with a

- 5 given determination is taken as indicating the need for retesting.

09618178 071800

## APPENDIX A

```

/* The probability distributions in Figures 4, 5, 6, and 7, respectively,
   correspond to the values of xx_prob, xy_prob, yy_prob, and ns_prob, for
   all possible values of the preprocessed reaction values (x_val and y_val)
   in the range of interest (0.0 to 3.0). */

```

```

/* We assume that the following global variables are set... */
double x_pos_mean, x_neg_mean, y_pos_mean, y_neg_mean;
double x_val, y_val;

```

```

/* And we set the following globals... */
double xx_prob, xy_prob, yy_prob, ns_prob;

```

```

#define POS_VARIANCE 0.25
#define POS_VARIANCE_INCREMENT 0.00
#define NEG_VARIANCE 0.05
#define NEG_VARIANCE_INCREMENT 0.10
#define HET_VARIANCE 0.10
#define HET_VARIANCE_INCREMENT 0.20

```

```

#define COND_NEG_PROB(val,given_val,val_mean) \
    normal_prob(val_mean-val,NEG_VARIANCE NEG_VARIANCE_INCREMENT*given_val)

```

```

#define COND_HET_PROB(val,given_val) \
    normal_prob(given_val-val,HET_VARIANCE + HET_VARIANCE_INCREMENT)

```

```

double normal_prob(deviation,sigma)
double deviation, sigma;
{
    double val=exp(-(deviation*deviation)/(2.0*sigma*sigma));
    return(val>=TINY_PROB ? val : TINY_PROB);
}

```

```

void compute_probs()
{
    double x_pos_prob, y_pos_prob, x_neg_prob, y_neg_prob;

    x_pos_prob=normal_prob((x_pos_mean-x_val), POS_VARIANCE);
    x_neg_prob=normal_prob((x_neg_mean-x_val), NEG_VARIANCE);
    y_pos_prob=normal_prob((y_pos_mean-y_val), POS_VARIANCE);
    y_neg_prob=normal_prob((y_neg_mean-y_val), NEG_VARIANCE);

    ns_prob=max(x_neg_prob * COND_NEG_PROB(y_val,x_val,y_neg_mean),
               y_neg_prob * COND_NEG_PROB(x_val,y_val,x_neg_mean));

    xx_prob=x_pos_prob * COND_NEG_PROB(y_val,x_val, y_neg_mean);

    yy_prob=y_pos_prob * COND_NEG_PROB(x_val,y_val, x_neg_mean);

    xy_prob= max(x_pos_prob * COND_HET_PROB(y_val,x_val),
               y_pos_prob * COND_HET_PROB(x_val,y_val));
}

```